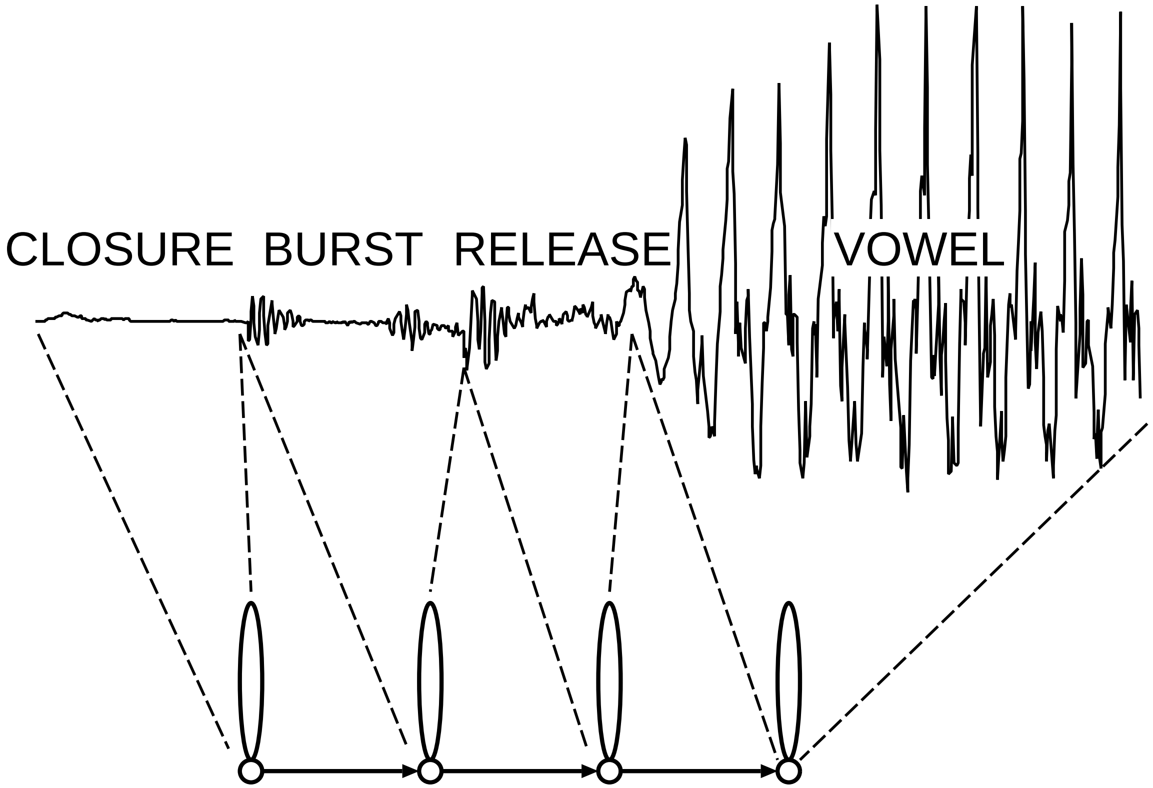
Cognitive Media Processing #8

Nobuaki Minematsu

Menu of the last lecture

- Fundamentals of statistical speech recognition
- Acoustic models for speech recognition
- From word models to subword models
- Speech recognition using grammars
- A small demo of automatic broadcast captioning
- Recommended books

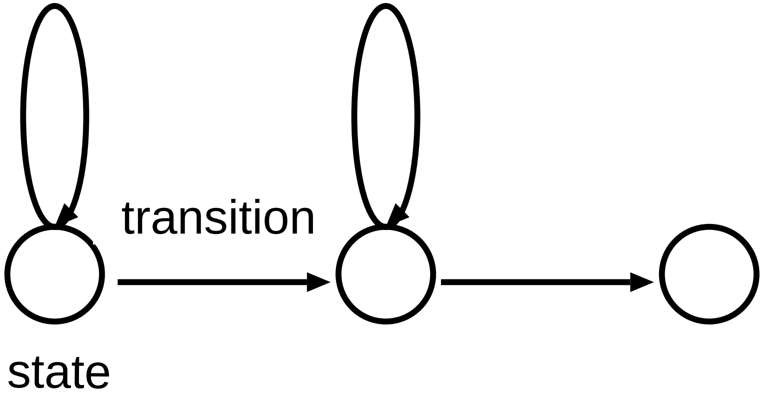
HMM as generative model



Probabilistic generative model

State transition is modeled as transition probability.
Output features are modeled as output probability.

Hidden Markov Process



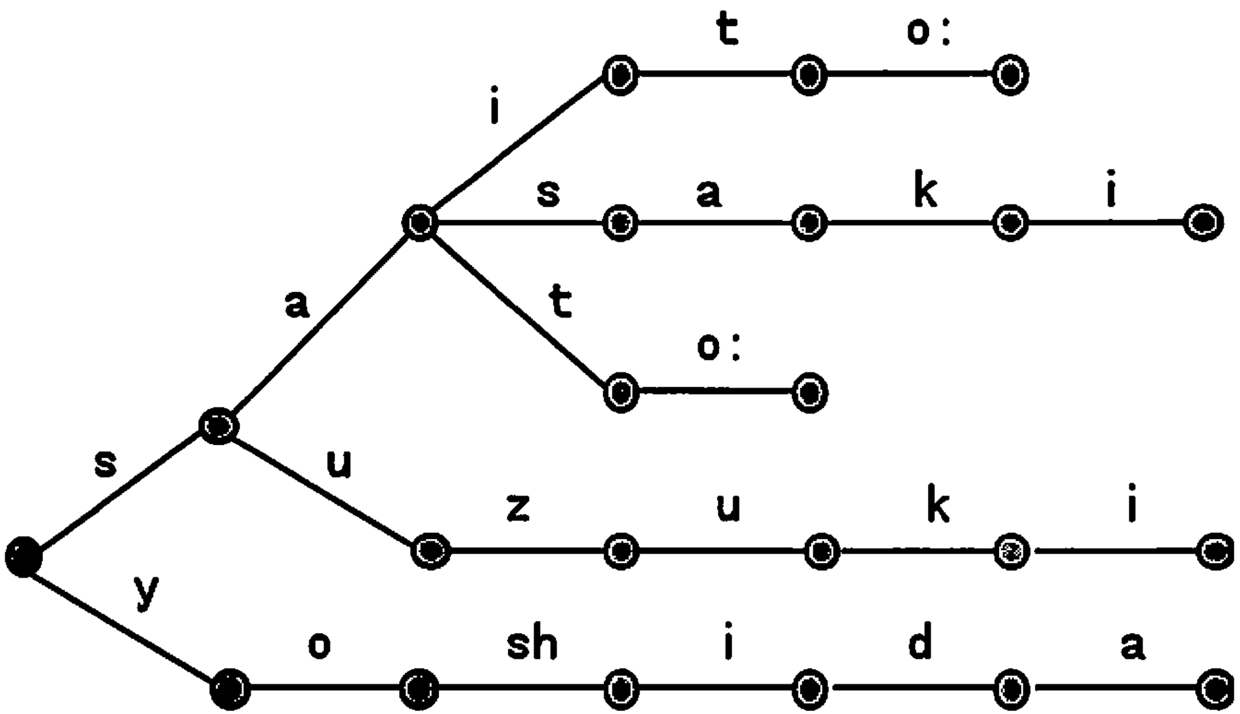
$$P(x_n|\ \underbrace{x_{n-1},\cdots,x_1}_{\text{previous observations}}\ ) = P(x_n|\ \underbrace{S_n}_{\text{current state}}\ )$$

Observation sequence : $x_1, x_2, \cdots, x_n, \cdots$

(Hidden) state sequence : $S_1, S_2, \cdots, S_n, \cdots$

- Previous observations cannot determine the current state uniquely.
- Signals (features) are observed but states are hidden.

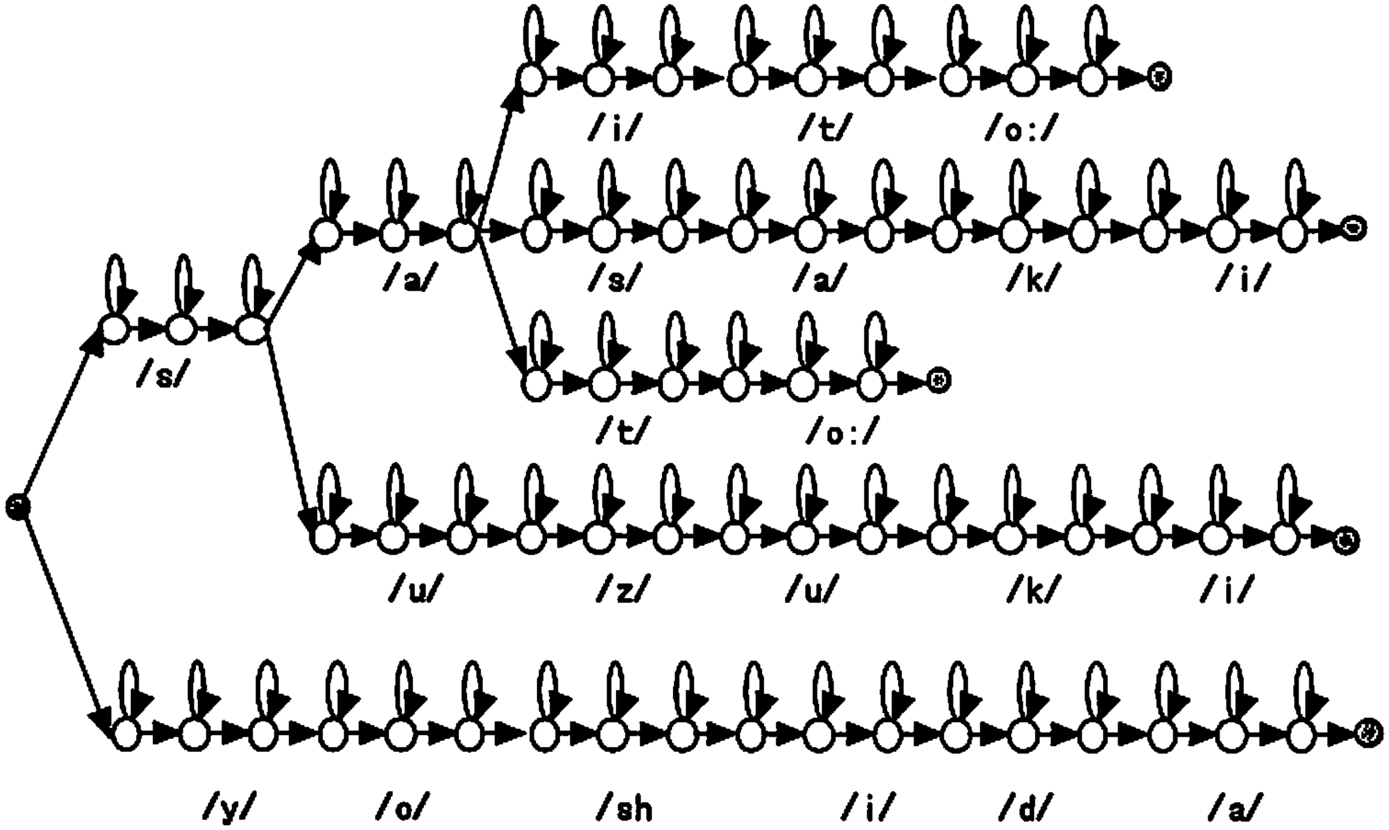
Tree lexicon (compact representation of the words)



The following words are stored as a tree.

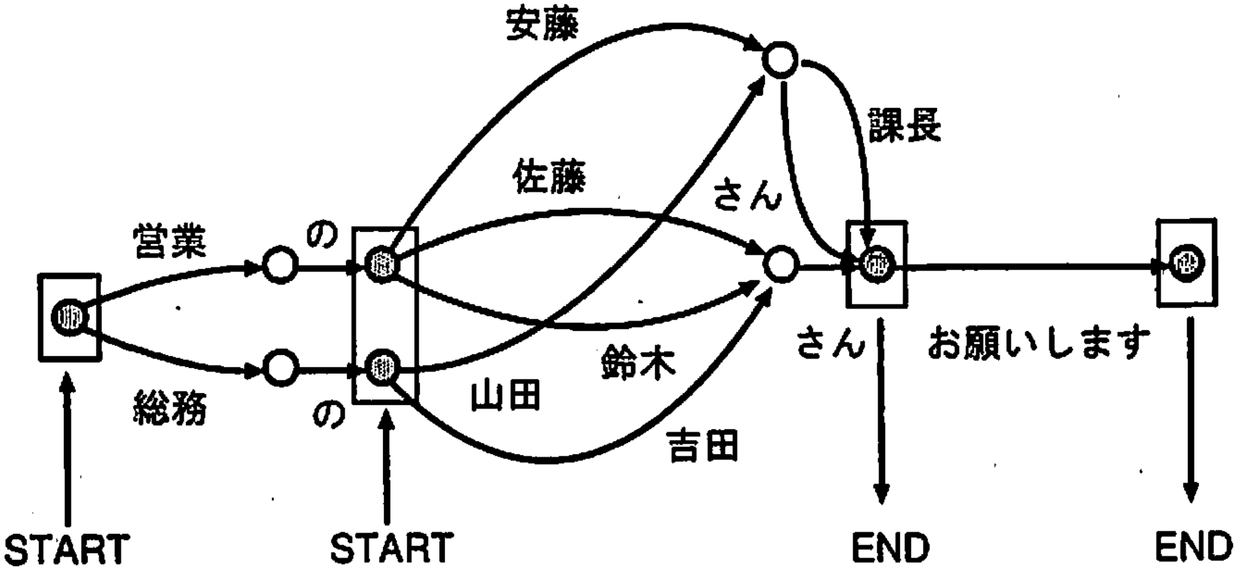
saito: (斉藤), sasaki (佐々木), sato: (佐藤)
suzuki (鈴木) , yoshida (吉田)

Tree-based lexicon using phoneme HMMs



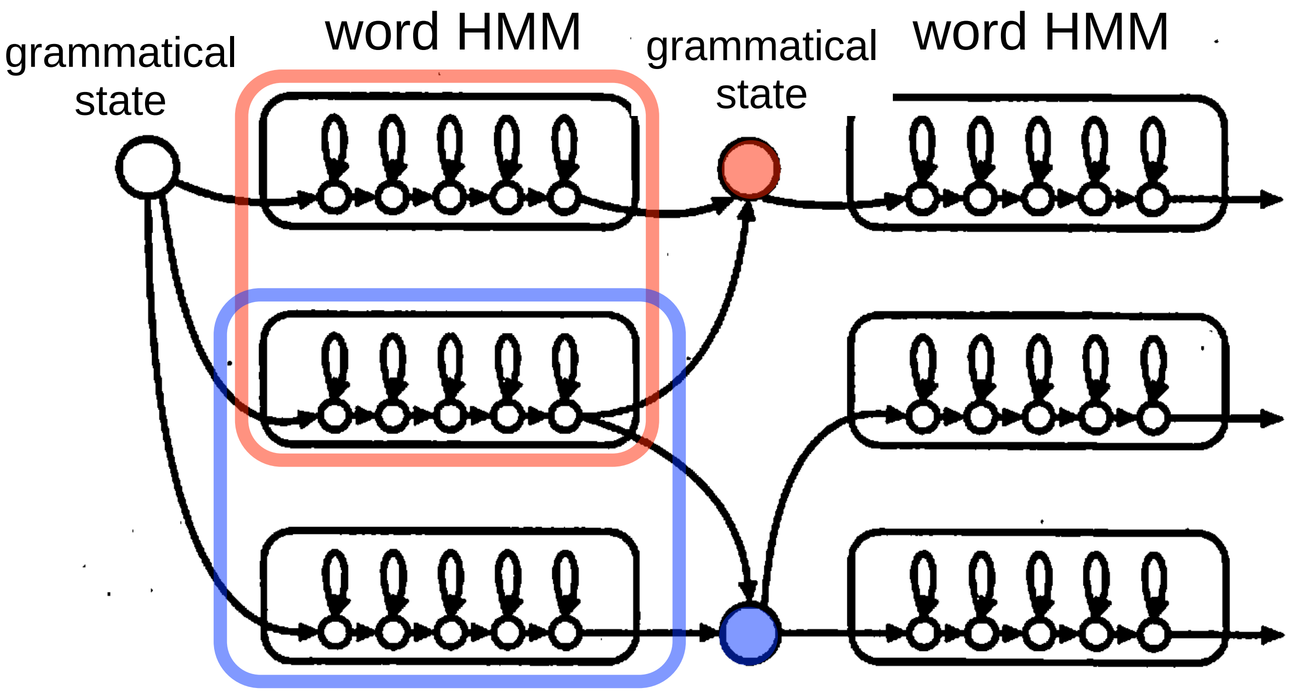
Generation of state-based network containing all the candidate words

Network grammar with a finite set of states



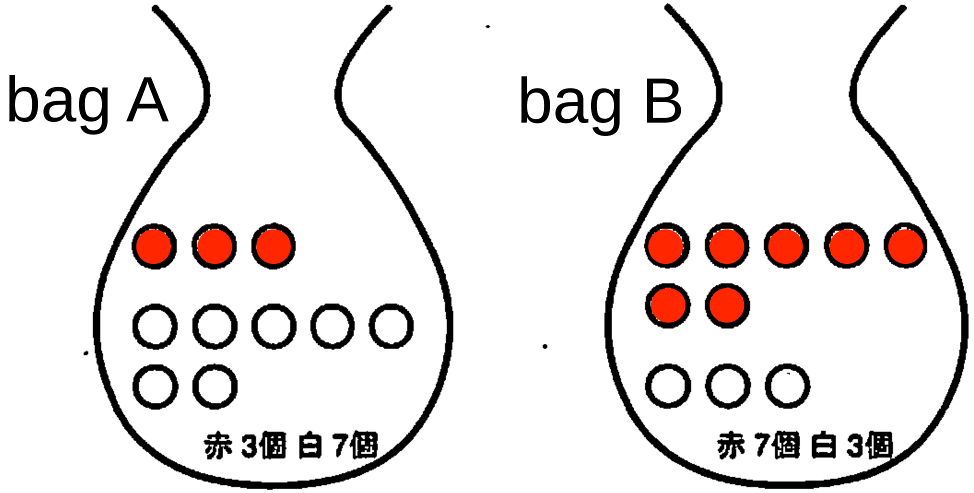
A sentence is accepted if it starts at one of the initial states and ends at one of the final states.

Speech recognition using a network grammar



When a grammatical state has more than one preceding words, the word of the maximum probability (or words with higher probabilities) is adopted and it will be connected to the following candidate words.

Probabilistic decision



Observation: You pick a ball three times. The colors are ● ○ ●.

Probabilities of P(●○●| A) and P(●○●| B)

$$袋A : \frac{3}{10}\times\frac{7}{10}\times\frac{3}{10}=0.063 \quad 袋B : \frac{7}{10}\times\frac{3}{10}\times\frac{7}{10}=0.147$$

Decision: The bag used is supposed to be B.

N-gram language model

The most widely-used implementation of P(w)

Only the previous N-1 words are used to predict the following word.
(N-1)-order Markov process

$$P(x_1,\cdots,x_n) = \underbrace{P(x_n|x_1,\cdots,x_{n-1})}_{\approx P(x_n|x_{n-N+1},\cdots,x_{n-1})} P(x_1,\cdots,x_{n-1})$$

$$\approx P(x_n|x_{n-N+1},\cdots,x_{n-1})P(x_1,\cdots,x_{n-1})$$

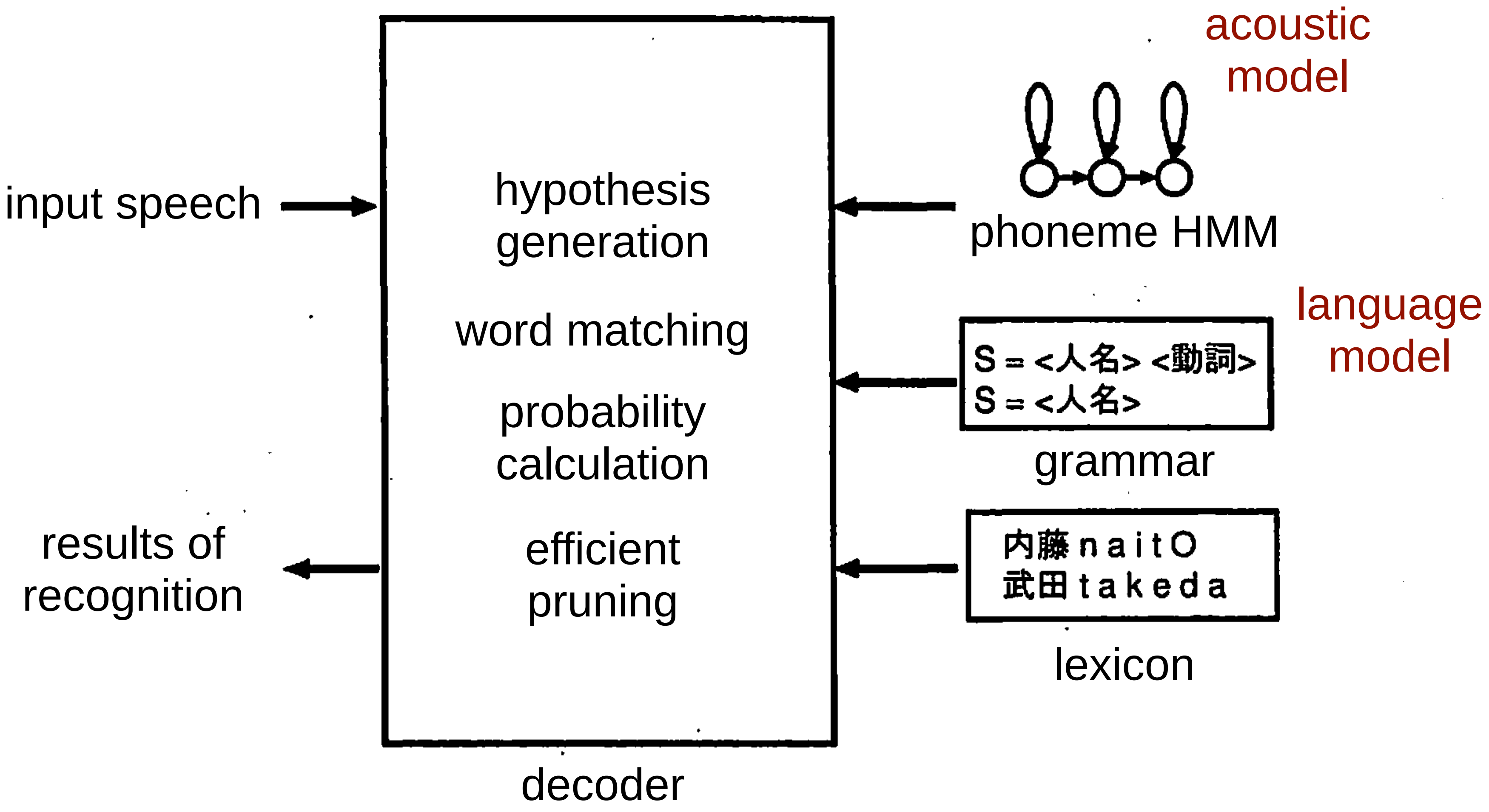
$$\approx \prod_{i=1}^{n} P(x_i|x_{n-N+1},\cdots,x_{i-1})$$

N-1 = 1 --> bi-gram
N-1 = 2 --> tri-gram

I’m giving a lecture on speech recognition technology to university students.

P(a | I’m, giving), P(lecture | giving, a), P(on | a, lecture),
P(speech | lecture, on), P(recognition | on, speech), ...

Development of a speech recognition system



input speech

hypothesis generation

word matching

probability calculation

efficient pruning

decoder

results of recognition

acoustic model

phoneme HMM

language model

S = <人名> <動詞>
S = <人名>

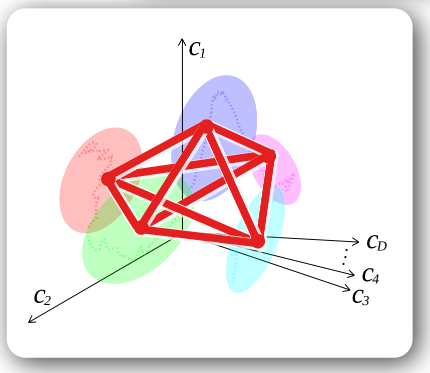
grammar

内藤 naitO
武田 takeda

lexicon

Title of each lecture

- Theme-1
  - ~~Multimedia information and humans~~
  - ~~Multimedia information and interaction between humans and machines~~
  - ~~Multimedia information used in expressive and emotional processing~~
  - ~~A wonder of sensation - synesthesia -~~
- Theme-2
  - ~~Speech communication technology - articulatory & acoustic phonetics -~~
  - ~~Speech communication technology - speech analysis -~~
  - ~~Speech communication technology - speech recognition -~~
  - Speech communication technology - speech synthesis -
- Theme-3
  - A new framework for “human-like” speech machines #1
  - A new framework for “human-like” speech machines #2
  - A new framework for “human-like” speech machines #3
  - A new framework for “human-like” speech machines #4



Speech Communication Tech.

- Speech synthesis -

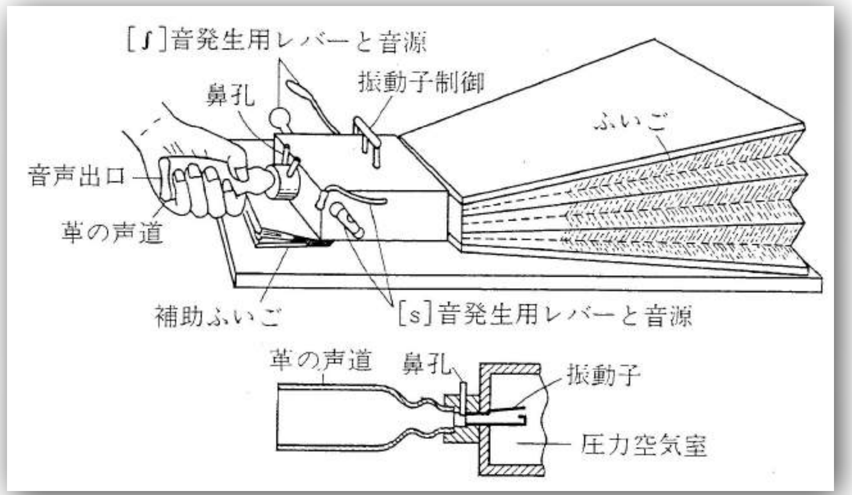
Nobuaki Minematsu

Today’s menu

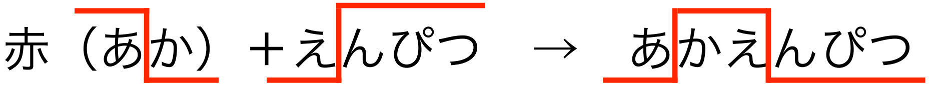
- Overview of text-to-speech conversion
  - From speaking machine to reading machine
- Text analysis
  - Text processing using units of sentences, phrases, and words
- Reading analysis
  - Assignment of reading (phonetic symbol + prosody) to each phoneme
- Waveform generation
  - Conversion of phonetic symbols + prosody to acoustic waveforms
- Some demos
  - Unit selection synthesis + HMM-based synthesis

The world oldest speech generator

- Speaking machine (Kempelen 1791)

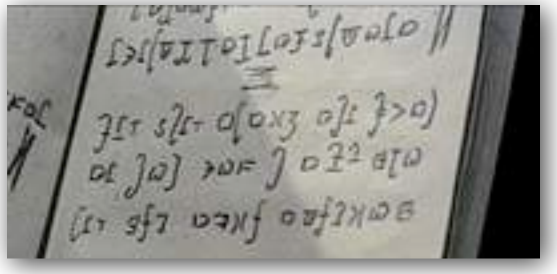


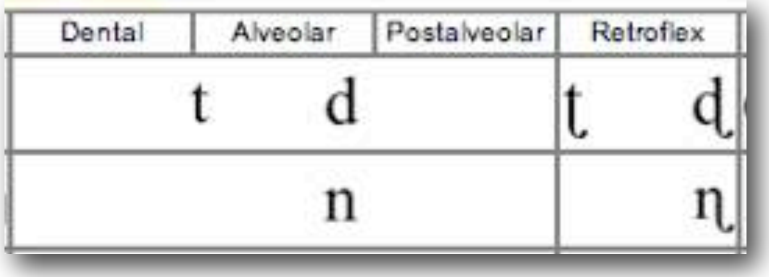
Difficulty of TTS

- Raw text alone is not sufficient “phonetically” for it to be read.
  - How to read Kanji? How to convert Kanji to Hiragana?
    - 今日の午後は，生物学の授業に出ました。
    - 今日＝きょう？こんにち？　　生物＝せいぶつ？なまもの？
  - Hiragana is similar to phonemic representation. It is enough for TTS?
    - とんぼ，とんねる，どんぐり
    - すいか，たべますか？
  - If text is represented by phonetic symbols, is it enough?
    - How about prosodic features which are needed for text-to-speech conversion?
    - Intonation, word accent, durational control (speaking rate), etc.
    - 
      赤（あか）＋えんぴつ　→　あかえんぴつ
  - Only “read”-style speech? Only native speech?
    - Expressive (emotional) speech
    - Non-native speech

Phones and phonemes

- Phones
  - A phone is the minimal unit of speech of any language.
  - Phonetic symbols are language-independent and used by phoneticians to transcribe speech of any language. Defined by by Int. Phonetic Association.
  - Should be used like [a b c d e f g].



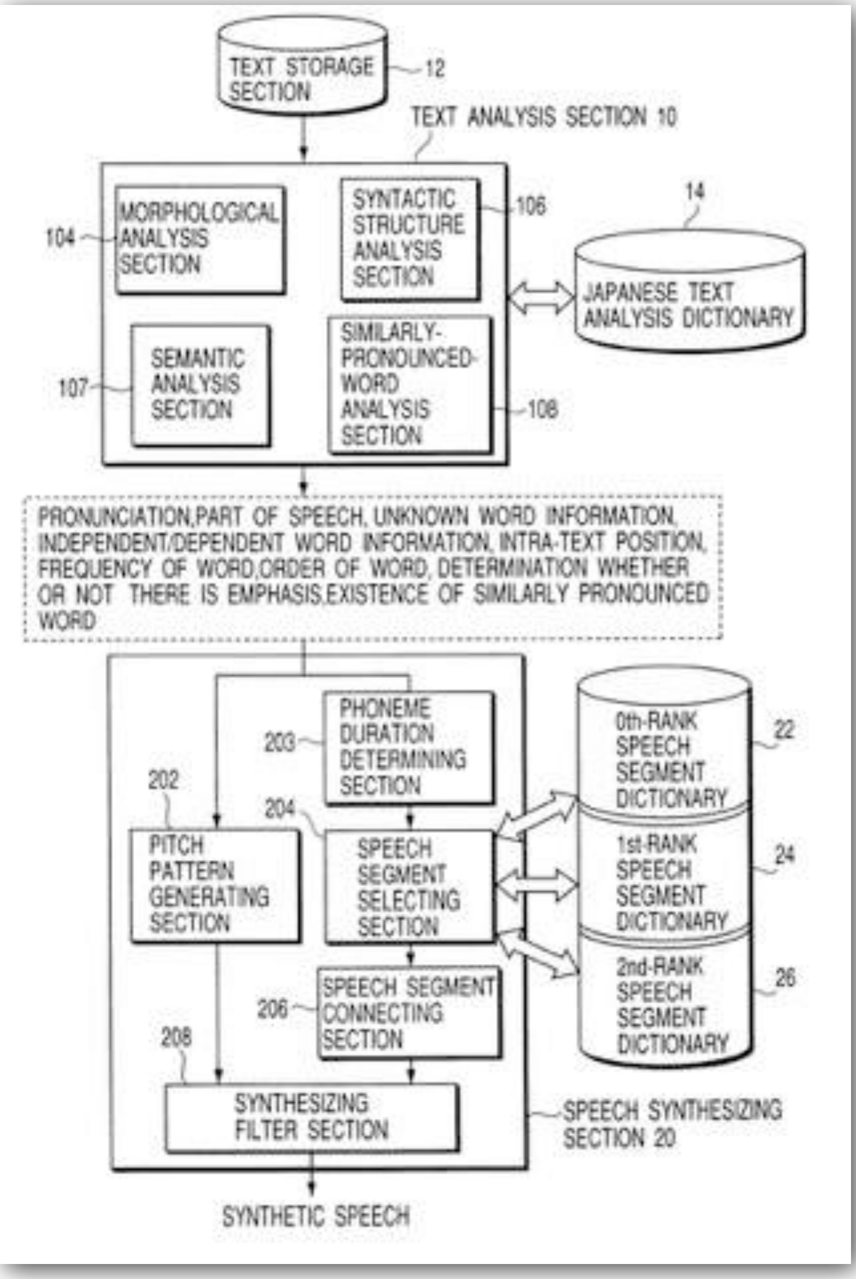


Dental	Alveolar	Postalveolar	Retroflex
	t d		ʈ ɖ
	n		ɳ

- Phonemes
  - A phoneme is the minimal unit of speech of a specific language, perceived by native speakers of that language.
  - Phonemic symbols are language-dependent and used by ordinary people to transcribe speech of that language. Can be defined by a user.
  - Should be used like / a b c d e f g /.

/arajurugeNzituo/ → [ɐɾɐʝ̞ɨɾɨ̠ge̞ẽ̞ɲ̟d̠͡ʑit͡sɨo̜ʔ]

Overview of text-to-speech conversion


TEXT STORAGE SECTION
12
TEXT ANALYSIS SECTION 10
104
MORPHOLOGICAL ANALYSIS SECTION
SYNTACTIC STRUCTURE ANALYSIS SECTION
106
14
JAPANESE TEXT ANALYSIS DICTIONARY
107
SEMANTIC ANALYSIS SECTION
SIMILARLY-PRONOUNCED-WORD ANALYSIS SECTION
108
PRONUNCIATION,PART OF SPEECH, UNKNOWN WORD INFORMATION, INDEPENDENT/DEPENDENT WORD INFORMATION, INTRA-TEXT POSITION, FREQUENCY OF WORD,ORDER OF WORD, DETERMINATION WHETHER OR NOT THERE IS EMPHASIS,EXISTENCE OF SIMILARLY PRONOUNCED WORD
203
PHONEME DURATION DETERMINING SECTION
202
204
PITCH PATTERN GENERATING SECTION
SPEECH SEGMENT SELECTING SECTION
0th-RANK SPEECH SEGMENT DICTIONARY
22
1st-RANK SPEECH SEGMENT DICTIONARY
24
2nd-RANK SPEECH SEGMENT DICTIONARY
26
206
SPEECH SEGMENT CONNECTING SECTION
208
SYNTHESIZING FILTER SECTION
SPEECH SYNTHESIZING SECTION 20
SYNTHETIC SPEECH

Conversion of any text input to its reading

今日の午後は，生物学の授業に出ました。

Conversion of reading to waveforms

今日＝キョー？コンニチ？

生物＝セイブツ？ナマモノ？

キョーノゴゴワ／セイブツガクノジュギョーニ／デマシタ

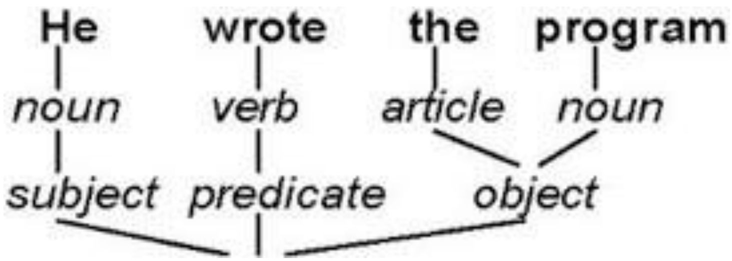
は＝ハ？ワ？

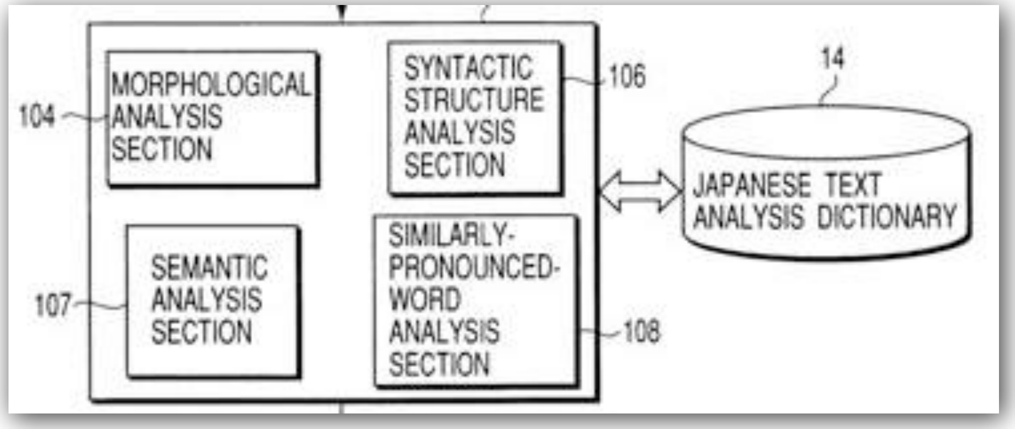
セイブツ＋ガク → セイブツガク

Vowel in シ of デマシタ＝voiced or unvoiced？

Text-to-speech synthesis

- Conversion from any input text to its sound sequence
  - In Japanese, Kanji have multiple ways of reading.
  - Kanji, Hiragana, Katakana, and Romaji
- Text analysis
  - Morphological analysis
    - An input sentence is divided into words (morphemes).
    - Part of speech (品詞) is assigned to each word.
  - Syntactic analysis
    - Syntactic structure in a sentence is extracted.
  - Semantic analysis
    - In many cases, it refers to correlation and co-occurrence among words.
- Reading analysis
  - Phonemic aspect
  - Prosodic aspect
- Waveform generation




MORPHOLOGICAL ANALYSIS SECTION
104
SYNTACTIC STRUCTURE ANALYSIS SECTION
106
14
JAPANESE TEXT ANALYSIS DICTIONARY
SEMANTIC ANALYSIS SECTION
107
SIMILARLY-PRONOUNCED-WORD ANALYSIS SECTION
108

Reading analysis

- Phonemic aspect (segmental aspect of speech)
  - Text to phonemic (Hiragana) representation
    - 今日の午後は → ky o: n o g o g o w a （きょうのごごは）
  - Phonemic representation to phonetic representation
    - ん(N) → n, m, ng ?
    - Some vowels have to be unvoiced.
    - :
- Prosodic aspect (supra-segmental aspect of speech)
  - Word-level processing
    - Word accent, accent sandhi in compound words
  - Phrase-level processing
    - Accent sandhi in connected words of a phrase
  - Sentence-level processing
    - Emphasis, phrasing

Raw text to Hiragana (phonemes)

- Two different reading styles of Japanese
  - On(音/オン)-reading and Kun(訓/くん)-reading
    - 生きる（い），生える（は），生物（セイ，なま），
- は and へ
  - は in 私は：“w a”, not “h a”
  - へ in 大学へ：“e” not “h e”
- Lengthened vowels
  - 消耗（しょうもう）：sh o: m o:
  - 映画（えいが）：e: g a
  - 大阪（おおさか）：o: sa ka
- Geminate consonant sounds (especially in numerical expressions)
  - 一巻（いっかん，ikkan）k becomes long (long consonant).
- Euphonic change of an unvoiced consonant to its voiced version (連濁)
  - 江戸川（えど＋かわ→えどがわ）

Numerical expressions

- Two different ways of reading numerical expressions
  - 03-5841-6662 : digit by digit
  - 123,456 yen : use of places, e.g. 12万3千4百5十6円
- Sound changes that are unique to numerical expressions
  - 523 is not ご　に　さん but ごおにいさん
- Geminate consonant sounds
  - 一本：いち＋ほん→いっぽん
  - １cm：いち＋せんち→いっせんち
- Euphonic change of unvoiced consonant to its voiced version (連濁)
  - 三本：さん＋ほん→さんぼん
  - 三階：さん＋かい→さんがい
  - 三回：さん＋かい→さんかい
- Exceptional cases
  - 一日：ついたち（いちにち is OK）
  - 二日：ふつか（ににち is OK）

Phoneme symbols to phonetic symbols

- Some vowels become unvoiced
  - If a vowel is surrounded by unvoiced consonants, it often become unvoiced.
    - アシカ（a sh i k a），エンピツ（e N p i ts u），スキヤキ（s u k i y a k i）etc
- Nasalized consonants
  - 株式会社（かぶしきがいしゃ，k a b u sh i k i g a i sh a)
- Syllabic nasal (撥音，ん)
  - 粘板岩（ねんばんがん）
    - [m] before p, b, and m
    - [ng] before k, g, and ng
    - [n] before t, d, n, and pause

Non-linguistic symbols and short forms

- Have to be converted into their phoneme sequences
  - %, kg, @
  - HMM, IT, IEEE
- Unknown words
  - Person’s names, city names (proper nouns)
  - Their reading have to be predicted using linguistic knowledge.
    - Grapheme-to-phoneme conversion

Reading analysis

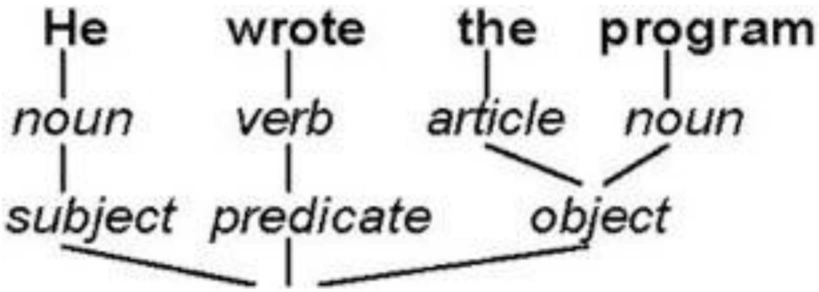
- Phonemic aspect (segmental aspect of speech)
  - Text to phonemic (Hiragana) representation
    - 今日の午後は → ky o: n o g o g o w a
  - Phonemic representation to phonetic representation
    - ん(N) → n, m, ng ?
    - Some vowels have to be unvoiced.
    - :
- Prosodic aspect (supra-segmental aspect of speech)
  - Word-level processing
    - Word accent, accent sandhi in compound words
  - Phrase-level processing
    - Accent sandhi in connected words of a phrase
  - Sentence-level processing
    - Emphasis, phrasing

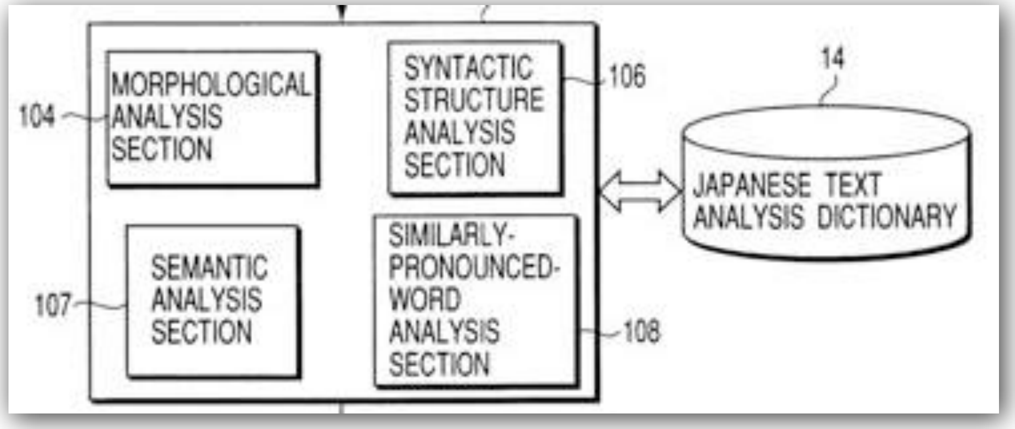
JEITA format

- Japan Electronics and Information Technology Industries Association
  - Text format designed specially for TTS input.
  - Proposed as standard and recommended by JEITA
  - Examples
    - 2006年の調査によると，日本全国で，約33%の家庭が，ペットを飼っているそうです。
    - ニセンロクネンノチョウサニヨルト，ニホンゼンコクデ，ヤクサンジュウサンパーセントノカテイガ，ペットヲカッテイルソウデス。
    - **ニセ'ンロクネンノ/チョ'ーサニヨルト_ニホ’ン_ゼ’ンコクデ_ヤ’ク/サ’ンジューサンパーセントノカテーガ_ペ’ットオ/カ’ッテイルソーデス%.**
    - ペットを家族のように考える人が増えたため，ペット関連の新しいビジネスも，生まれました。
    - ペットヲカゾクノヨウニカンガエルヒトガフエタタメ，ペットカンレンノアタラシイビジネスモ，ウマレマシタ。
    - **ペ'ットオ/カ'ゾクノヨーニカンガエルヒ%トガ_フ’エタタメ_ペットカ'ンレンノ_アタラシ'ー/ビ'ジネスモウマレマシ%タ.**

Text-to-speech synthesis

- Conversion from any input text to its sound sequence
  - In Japanese, multiple ways of reading have to be dealt with.
  - Kanji, Hiragana, Katakana, and Romaji
- Text analysis
  - Morphological analysis
    - An input sentence is divided into words (morphemes)
    - Part of speech (品詞) is assigned to each word.
  - Syntactic analysis
    - Syntactic structure in a sentence is extracted.
  - Semantic analysis
    - In many cases, it refers to correlation analysis of
    - words that occur frequently.
- Reading analysis
  - Phonemic aspect
  - Prosodic aspect
- Waveform generation




104
MORPHOLOGICAL ANALYSIS SECTION
SYNTACTIC STRUCTURE ANALYSIS SECTION
106
14
JAPANESE TEXT ANALYSIS DICTIONARY
107
SEMANTIC ANALYSIS SECTION
SIMILARLY-PRONOUNCED-WORD ANALYSIS SECTION
108

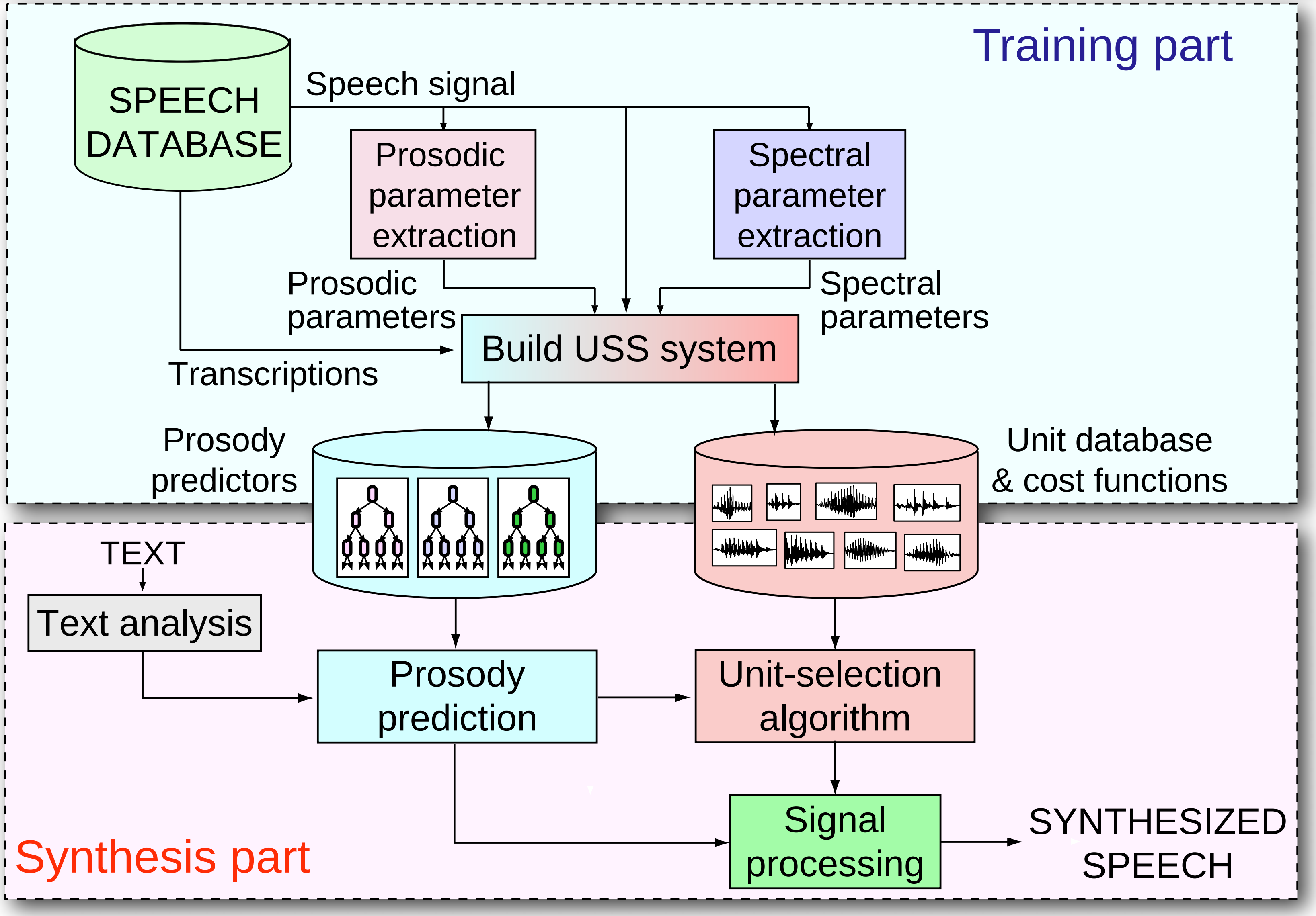
Two major methods of waveform generation

- Unit selection synthesis
  - A huge number of waveform units (templates) are stored in a database.
  - Adequate selection of waveform units is done based on input text.
  - The selected units are smoothly concatenated.
  - Additional pitch modification is often needed.

- HMM-based synthesis
  - A kind of “unit selection” synthesis
    - Units are not waveform units but spectrum units
  - A huge number of context-dependent phoneme HMMs are stored in a database.
  - Adequate selection of HMMs is done based on input text.
  - The selected HMMs are concatenated.
  - Pitch is realized by generating a source signals based on the desired pitch pattern.

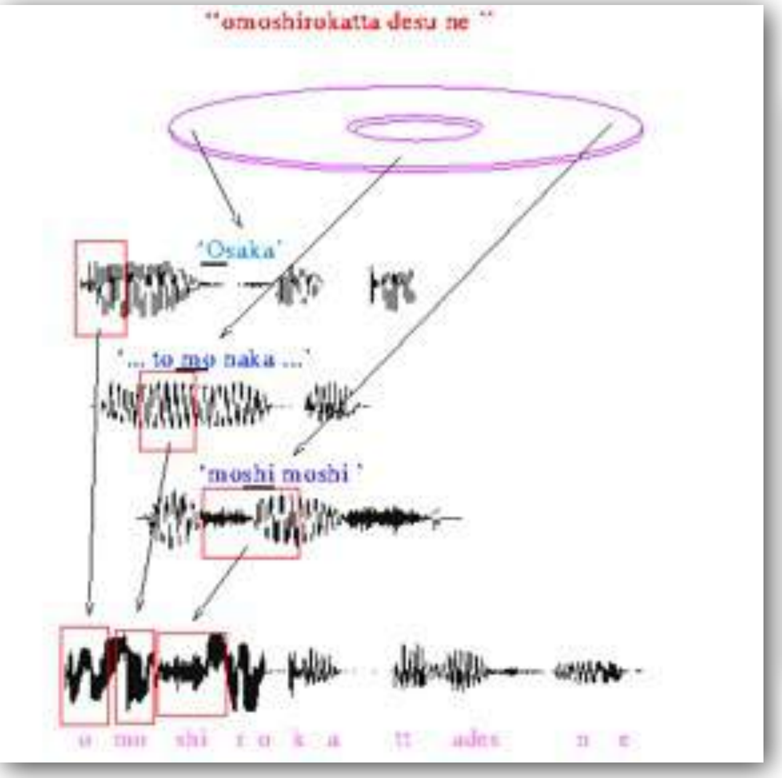
A diagram of unit selection synthesis

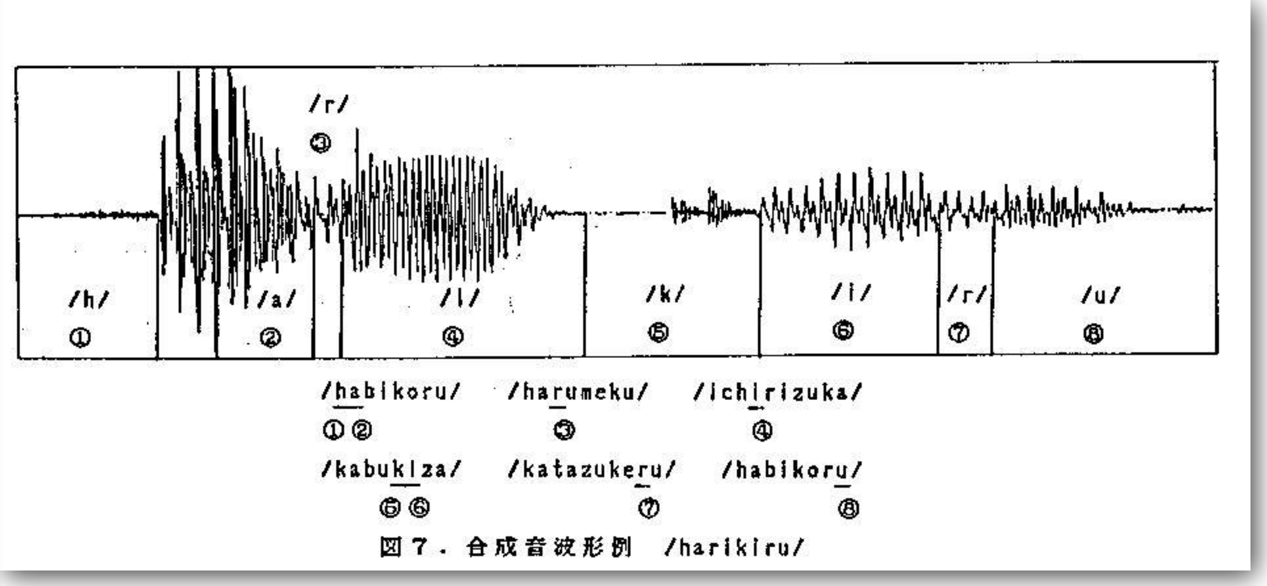
Unit-selection synthesis (USS) (1)



Unit selection synthesis

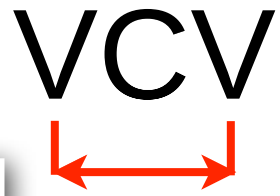
- A speech corpus is segmented using some linguistic units.
  - Phoneme, syllable, or in-between?
- Adequate selection of waveform units is done based on text input.
  - How to select adequate units?
- The selected units are smoothly concatenated.
  - If waveform units of the desired pitch level are not found, how to prepare the units?

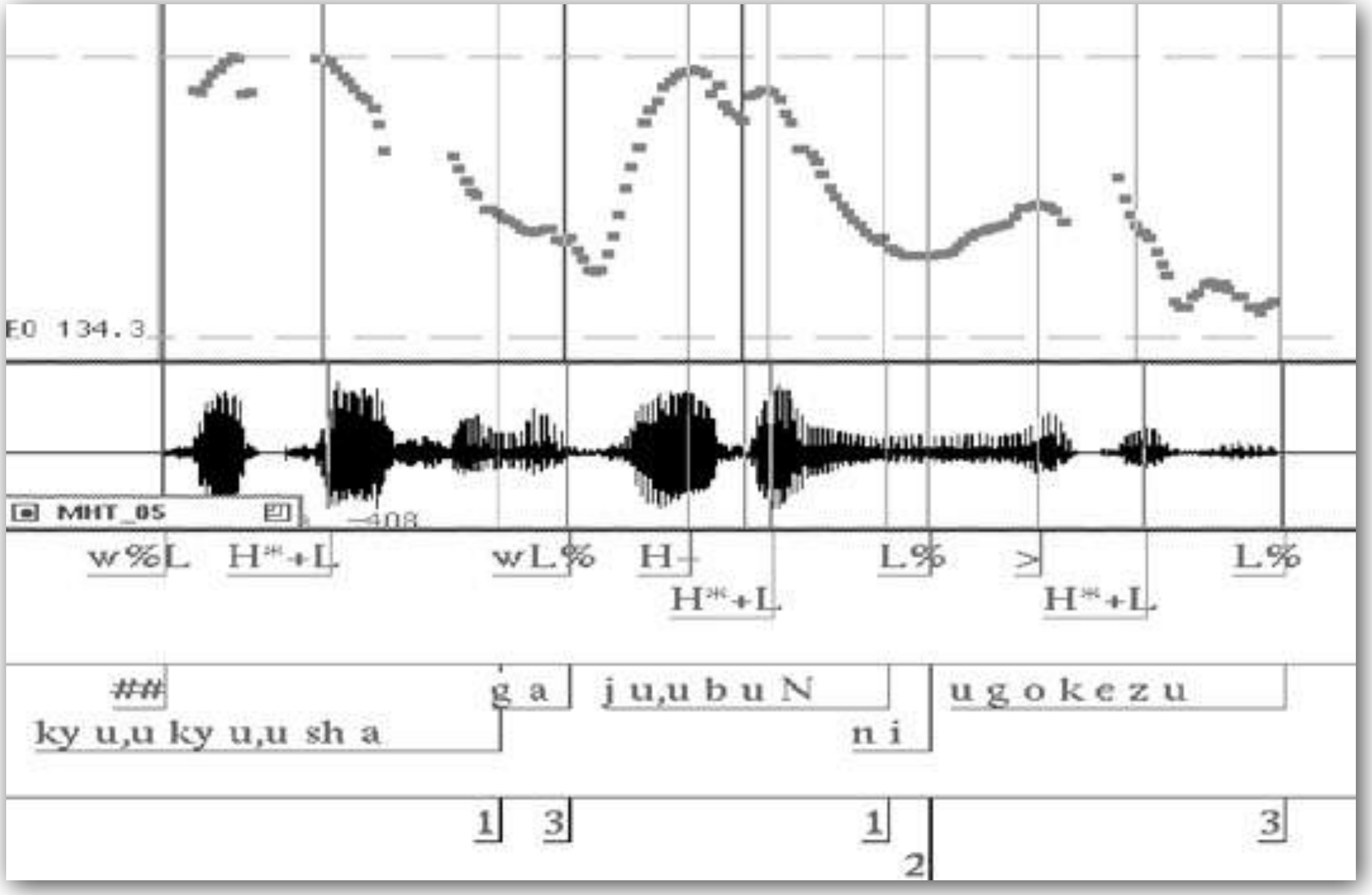




Unit selection synthesis

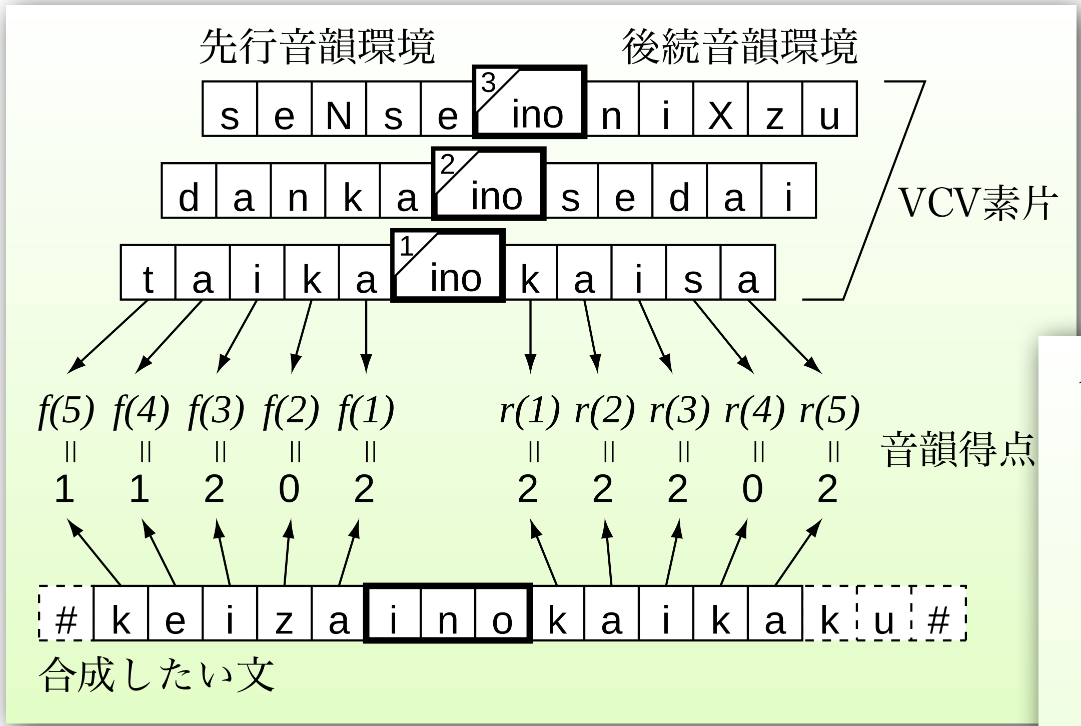
- A speech corpus is segmented using some linguistic units.
  - Phoneme, syllable, and VCV units
  - Prosodic attributes are also considered (added) in labeling.

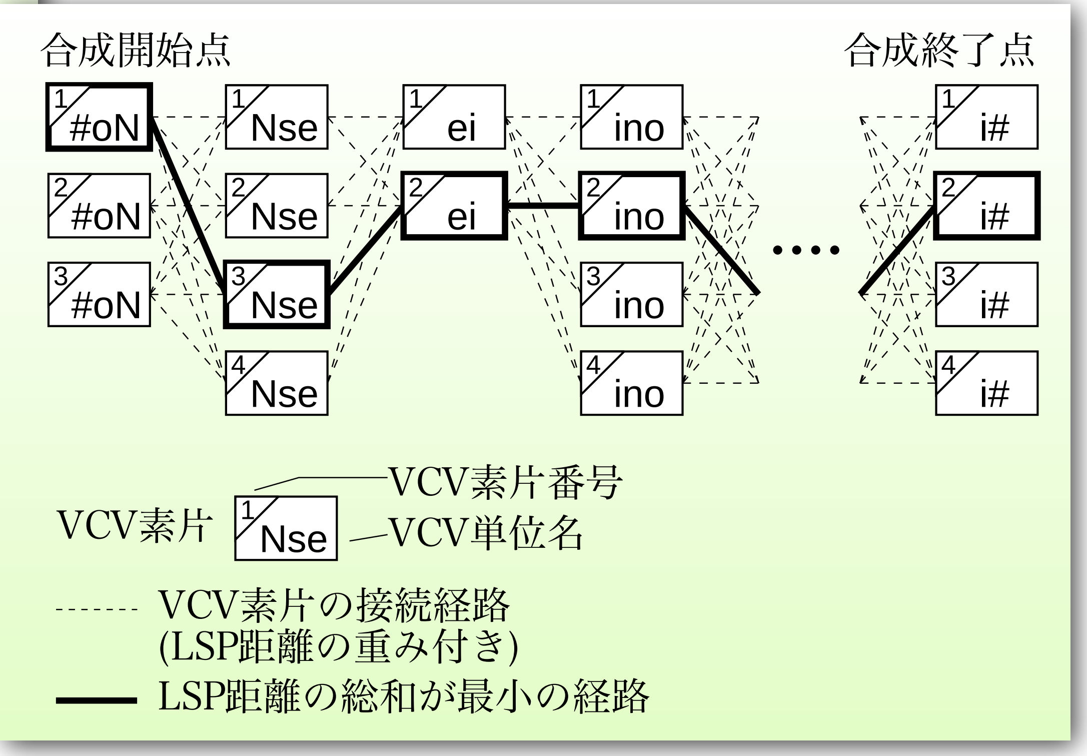




Unit selection synthesis

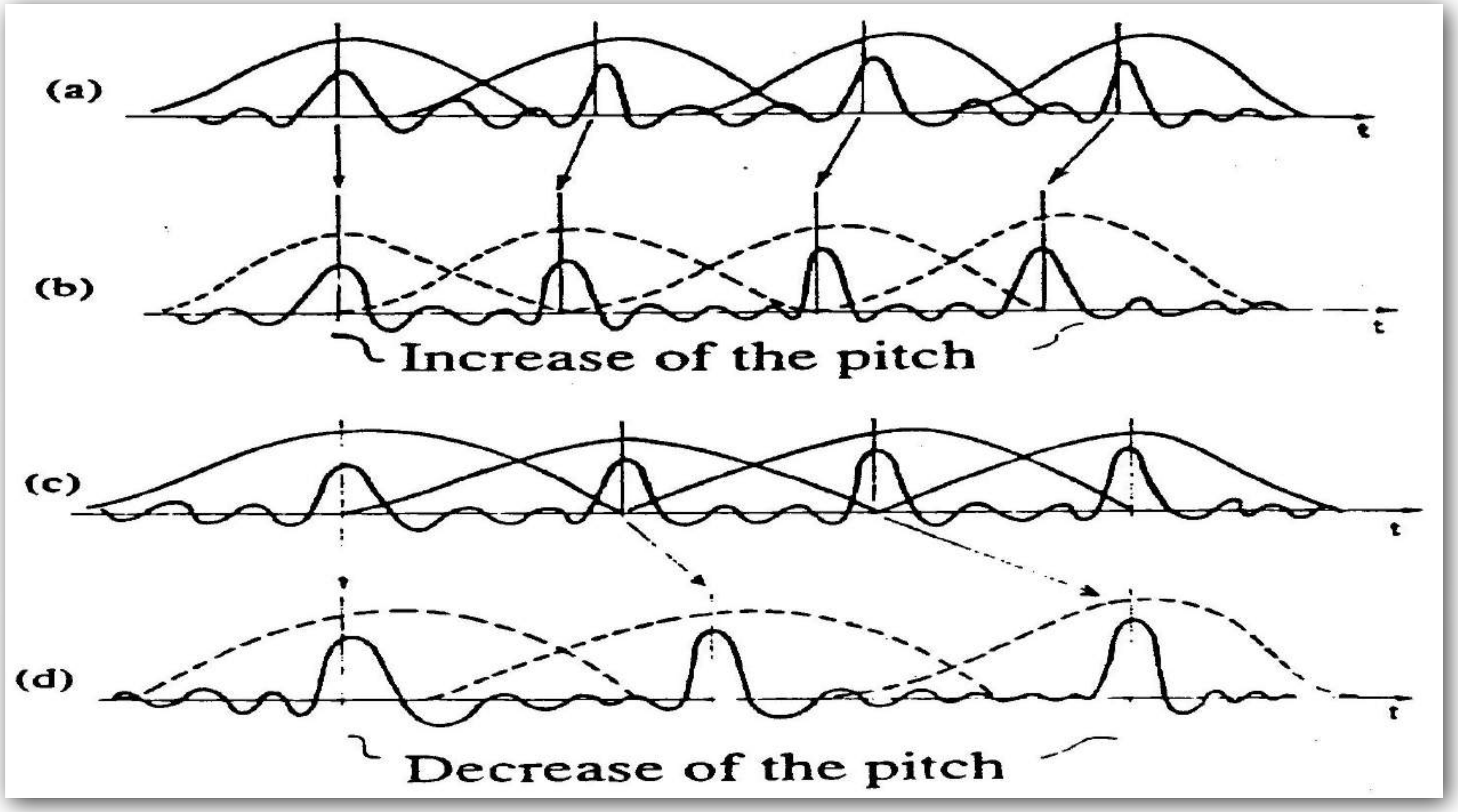
- Adequate selection of waveform units is done based on text input.
  - A cost function is defined and the unit that can minimize the cost is selected.





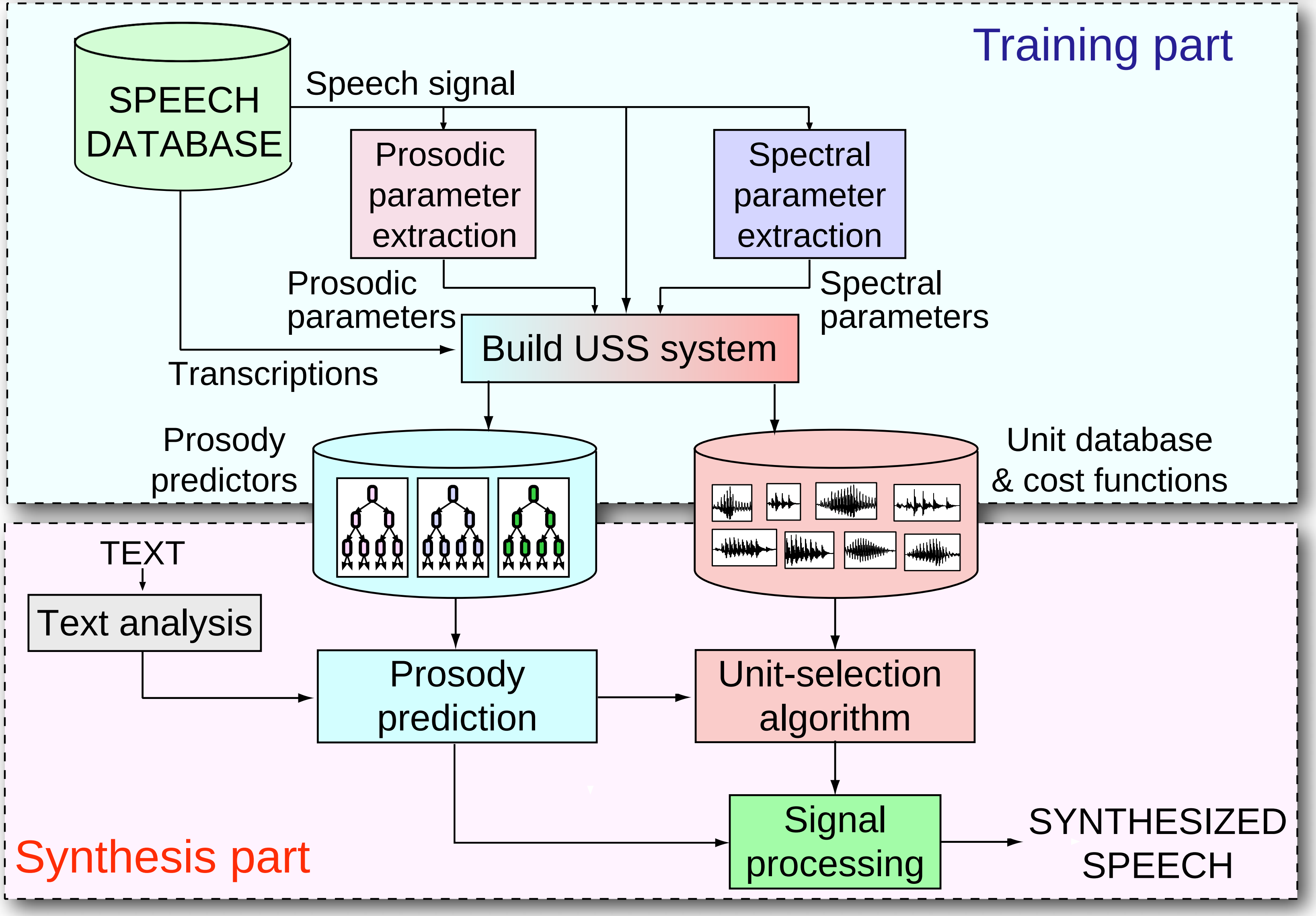
Unit selection synthesis

- The selected units are smoothly concatenated.
  - Pitch change is realized by PSOLA (Pitch Synchronous OverLap and Add)
  - Pitch waveforms are concatenated with overlap using required intervals.

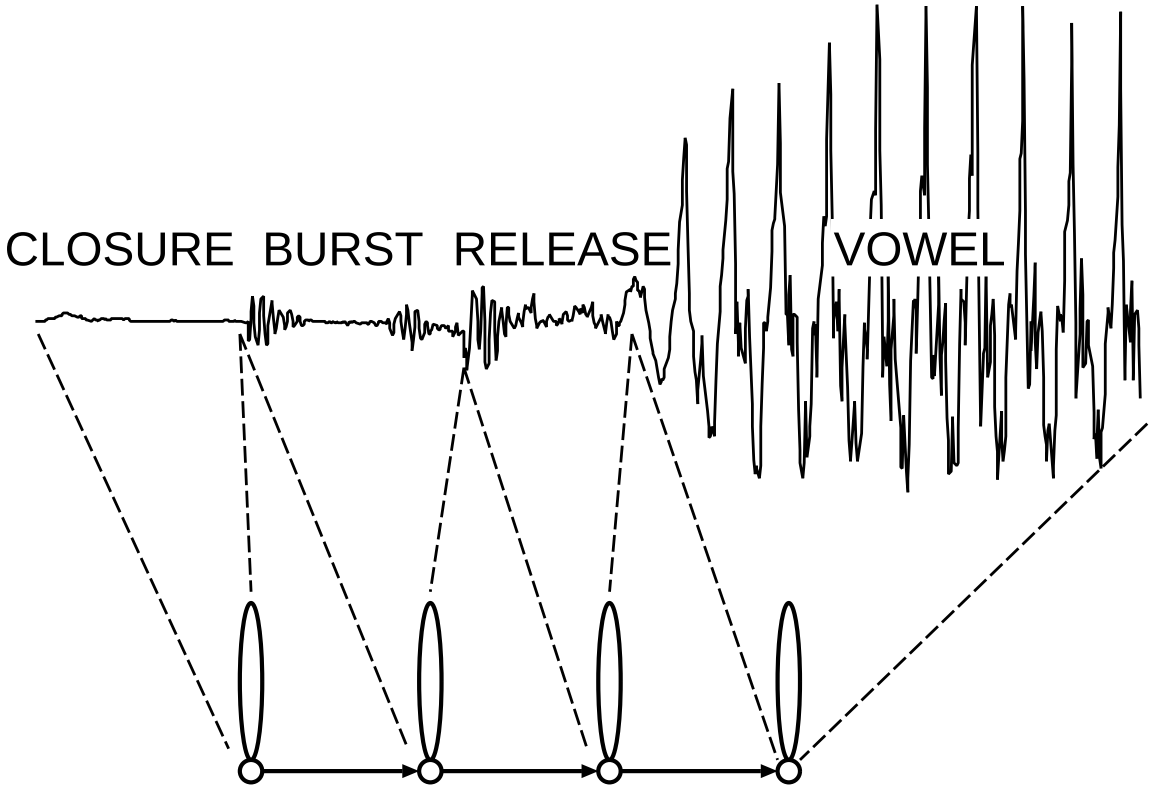


A diagram of unit selection synthesis

Unit-selection synthesis (USS) (1)



HMM as generative model

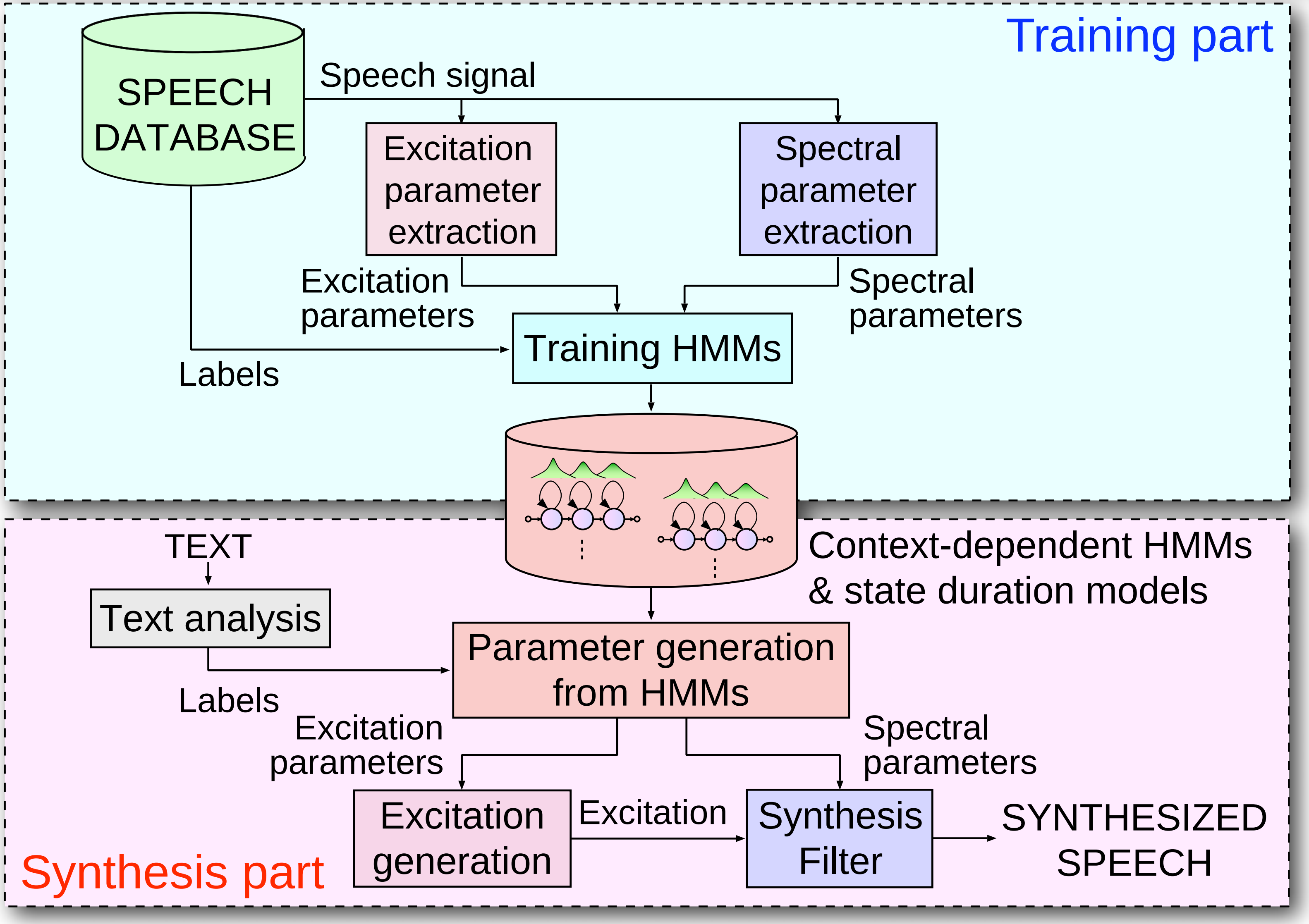


Probabilistic generative model

State transition is modeled as transition probability.
Output features are modeled as output probability.

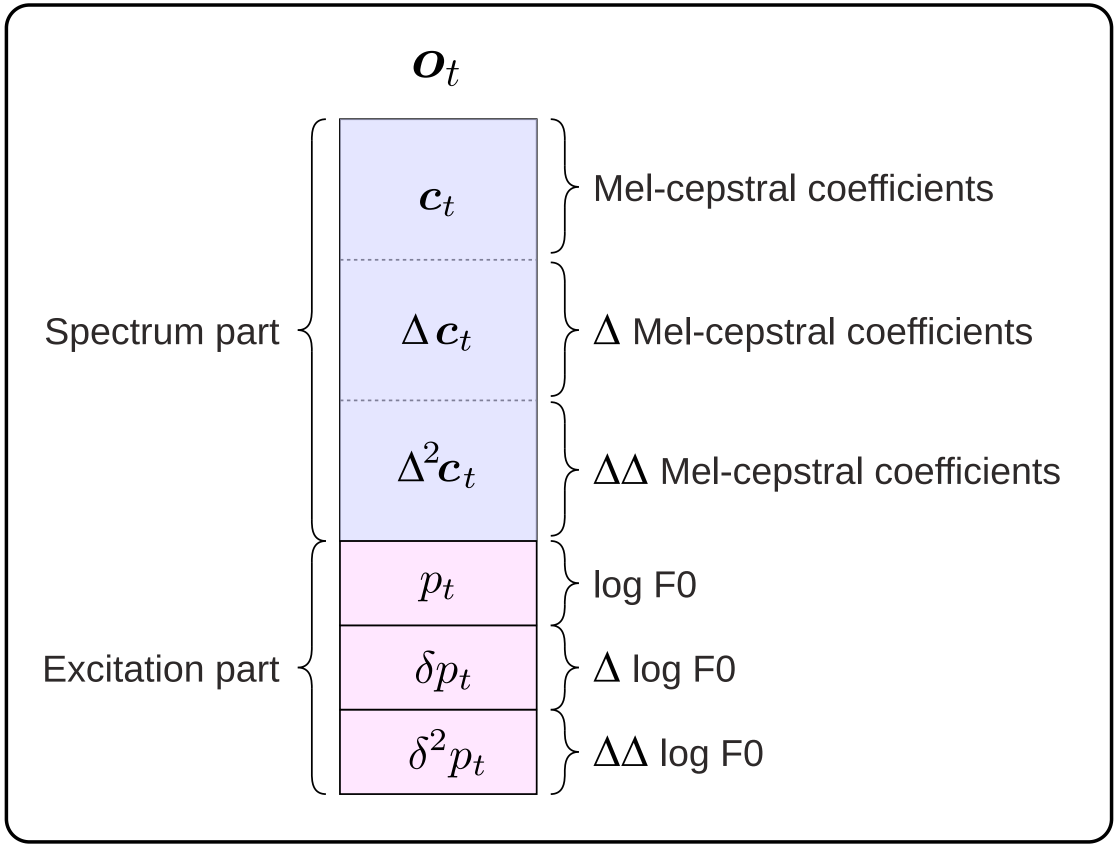
A diagram of HMM-based synthesis

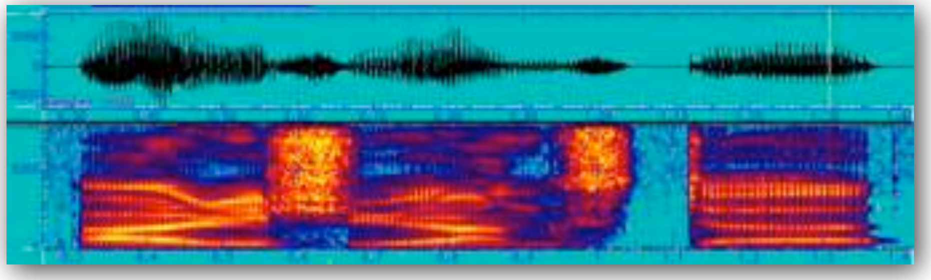
HMM-based speech synthesis system (HTS)



HMM-based synthesis

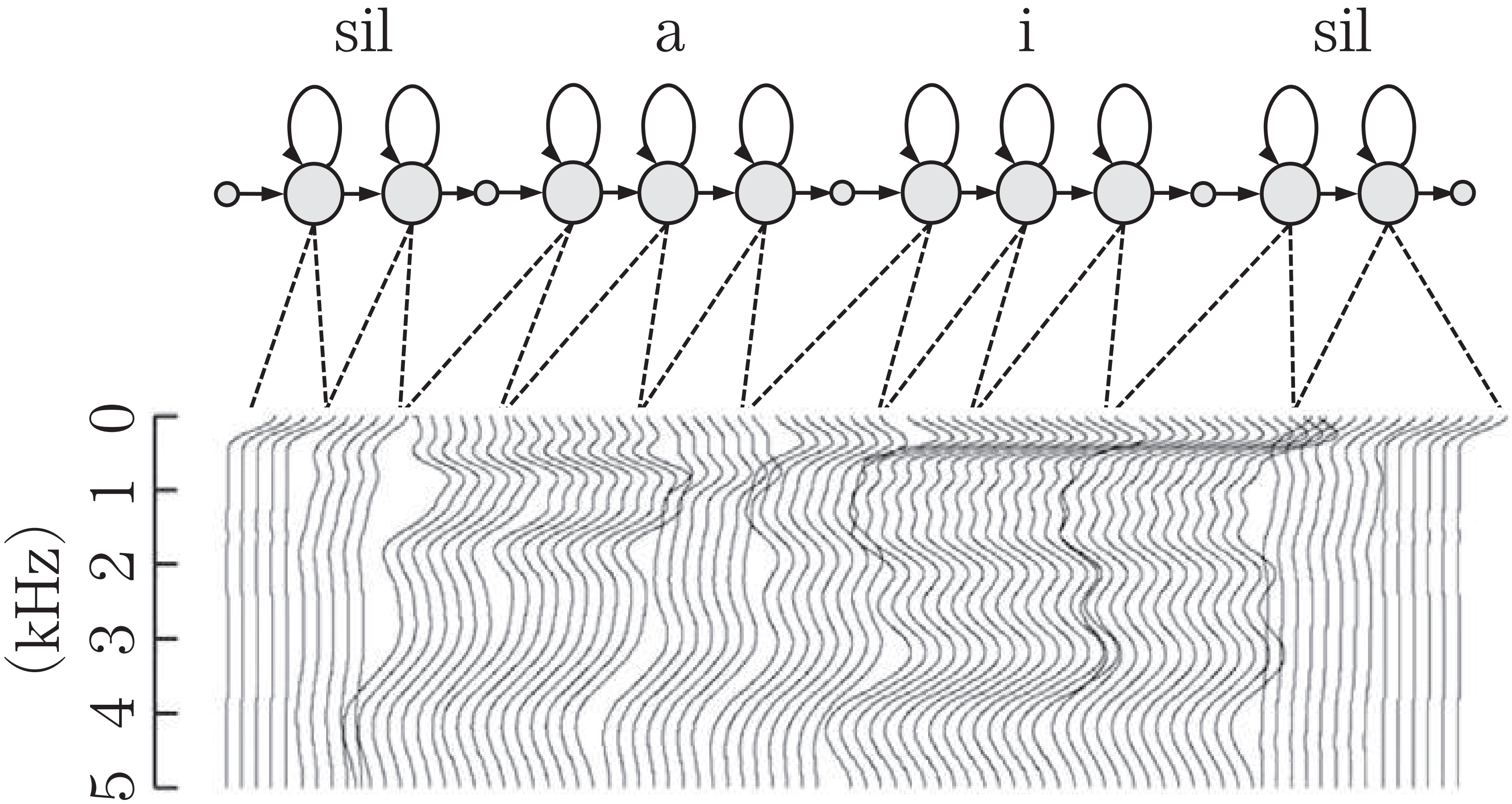
- Text -> HMM seq. -> most likely state seq. -> most likely spectrum seq.
- Spectrum seq. = adaptive filter
- By inputing source signals to the filter, waveforms can be obtained.





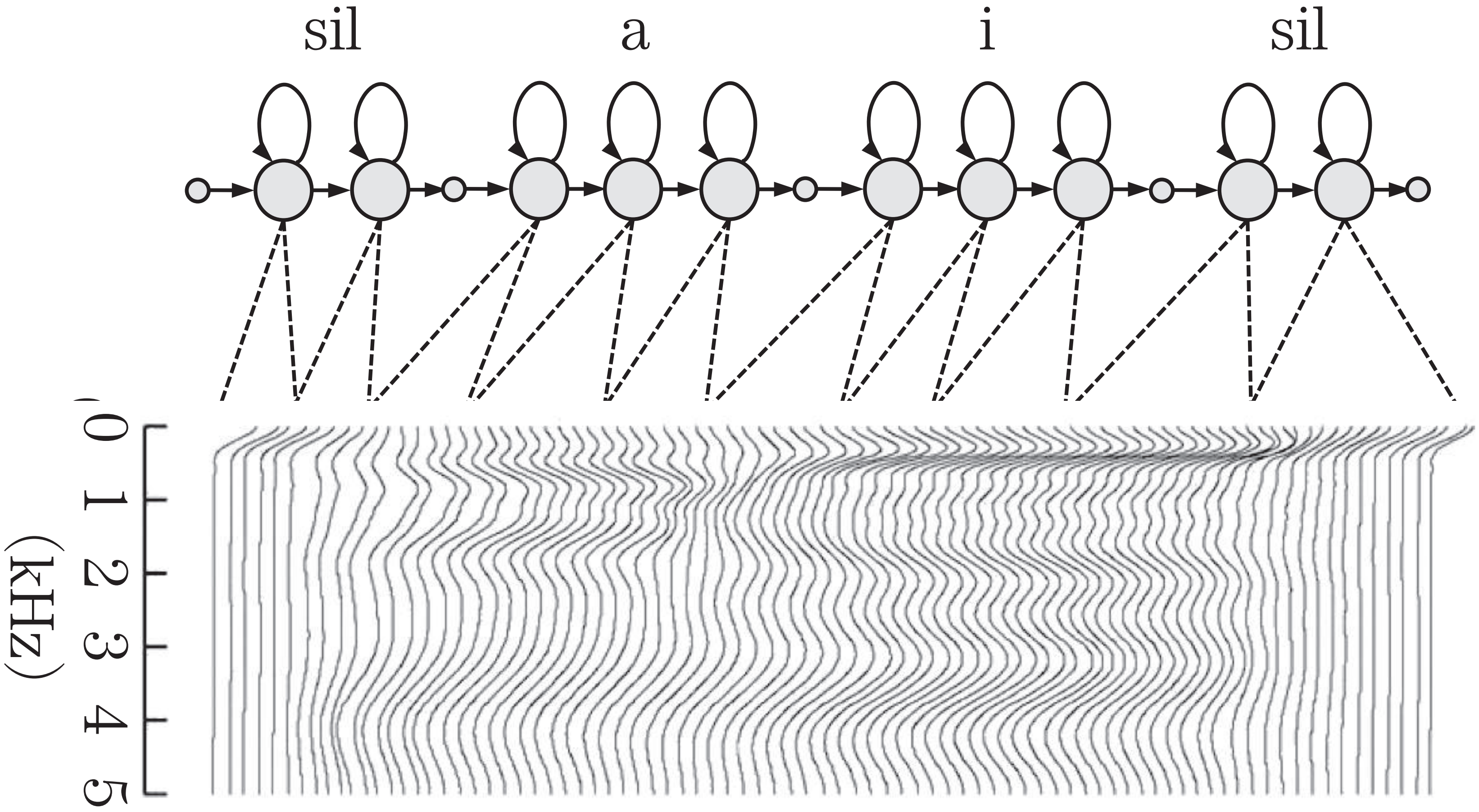
Spectrum generated from HMMs

- Text -> HMM seq. -> most likely state seq. -> most likely spectrum seq.
  - The most likely spectrum from a state = mean vector (spectrum) of the state
    --> the spectrum sequence has to have stepwise abrupt changes.



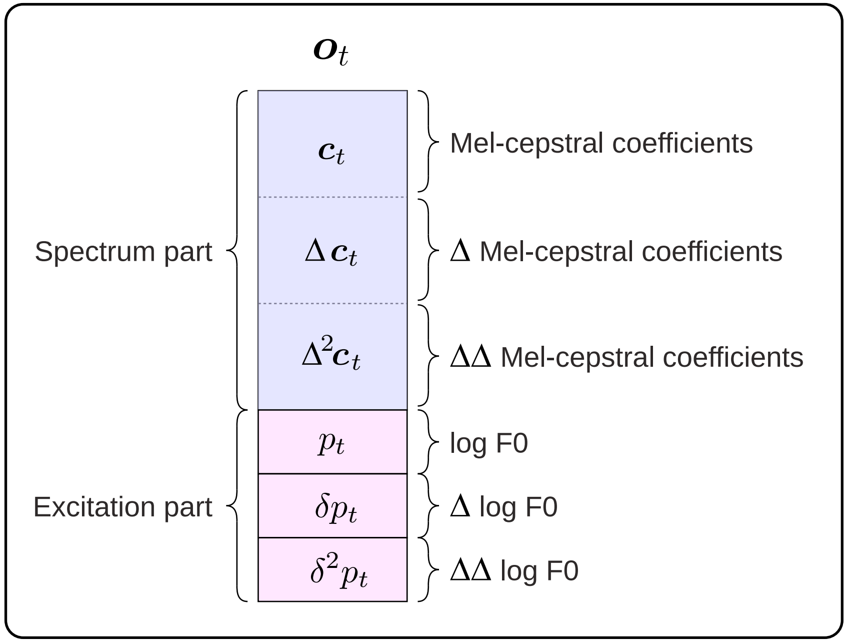
Spectrum generated from HMMs

- Text -> HMM seq. -> most likely state seq. -> most likely spectrum seq.
  - The most likely spectrum from a state = mean vector (spectrum) of the state
    --> the spectrum sequence has to have stepwise abrupt changes.



Static features and dynamic features

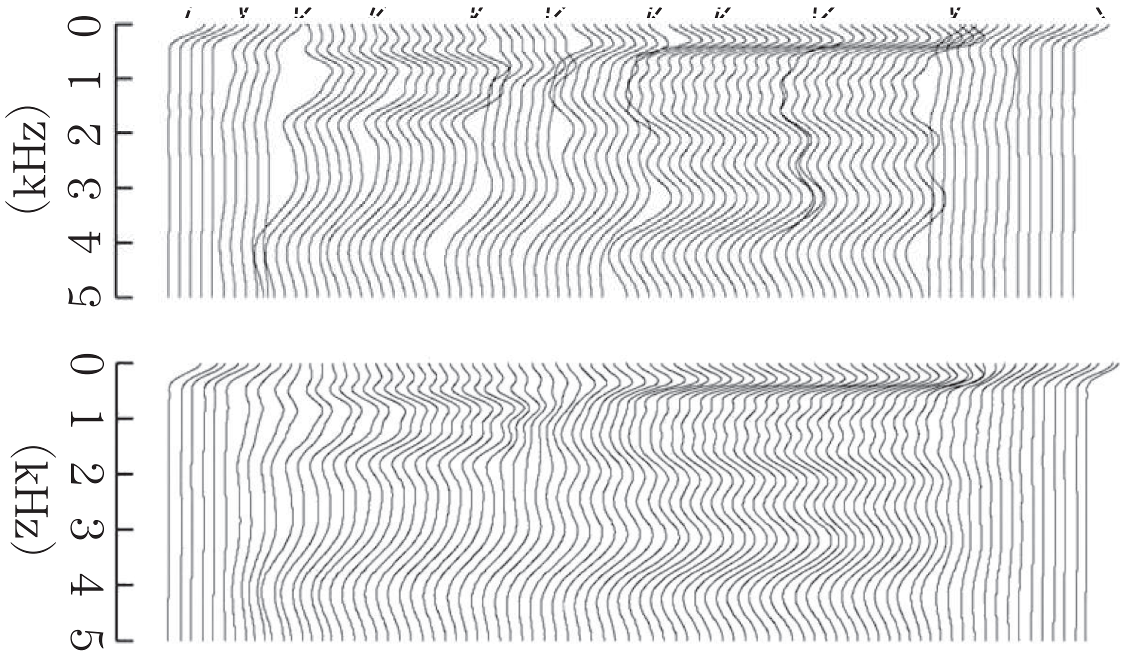
- Maximum likelihood generation of Cep sequences with constraints
  - ΔCep = velocity components, ΔΔCep = acceleration components
  - Cep + Δ Cep + ΔΔ Cep are used as features of HMM
  - What is needed is a sequence of Cep that is adequate for input text.
    - Cep sequence is generated by using Δ and Δ features as constraint.
    - Likelihood of ΔCep + ΔΔCep should be increased as well as that of Cep.



$\boldsymbol{o}_t$

Spectrum part: $\boldsymbol{c}_t$ Mel-cepstral coefficients; $\Delta\,\boldsymbol{c}_t$ Δ Mel-cepstral coefficients; $\Delta^2\boldsymbol{c}_t$ ΔΔ Mel-cepstral coefficients

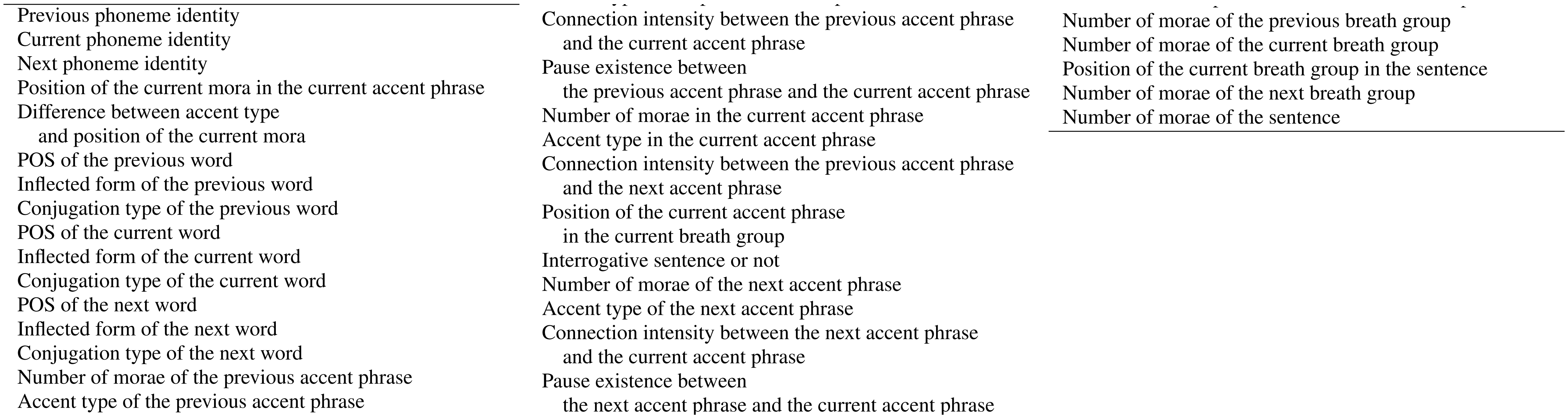
Excitation part: p_t log F0; δp_t Δ log F0; $\delta^2 p_t$ ΔΔ log F0



Context used for HMM for synthesis

- Context-dependent phoneme HMMs for ASR
  - Context = left phoneme and right phoneme
  - /a/ -> i-a+b, u-a+t, h-a+l, ...
  - #triphones = N x N x N (N : number of phonemes)
- Context-dependent phoneme HMMs for TTS
  - Context = left phoneme, right phoneme + so many linguistic factors
  - #context-dependent-phonemes = logically infinite !?
  - HMMs for TTS are by far much finer than HMMs for ASR.

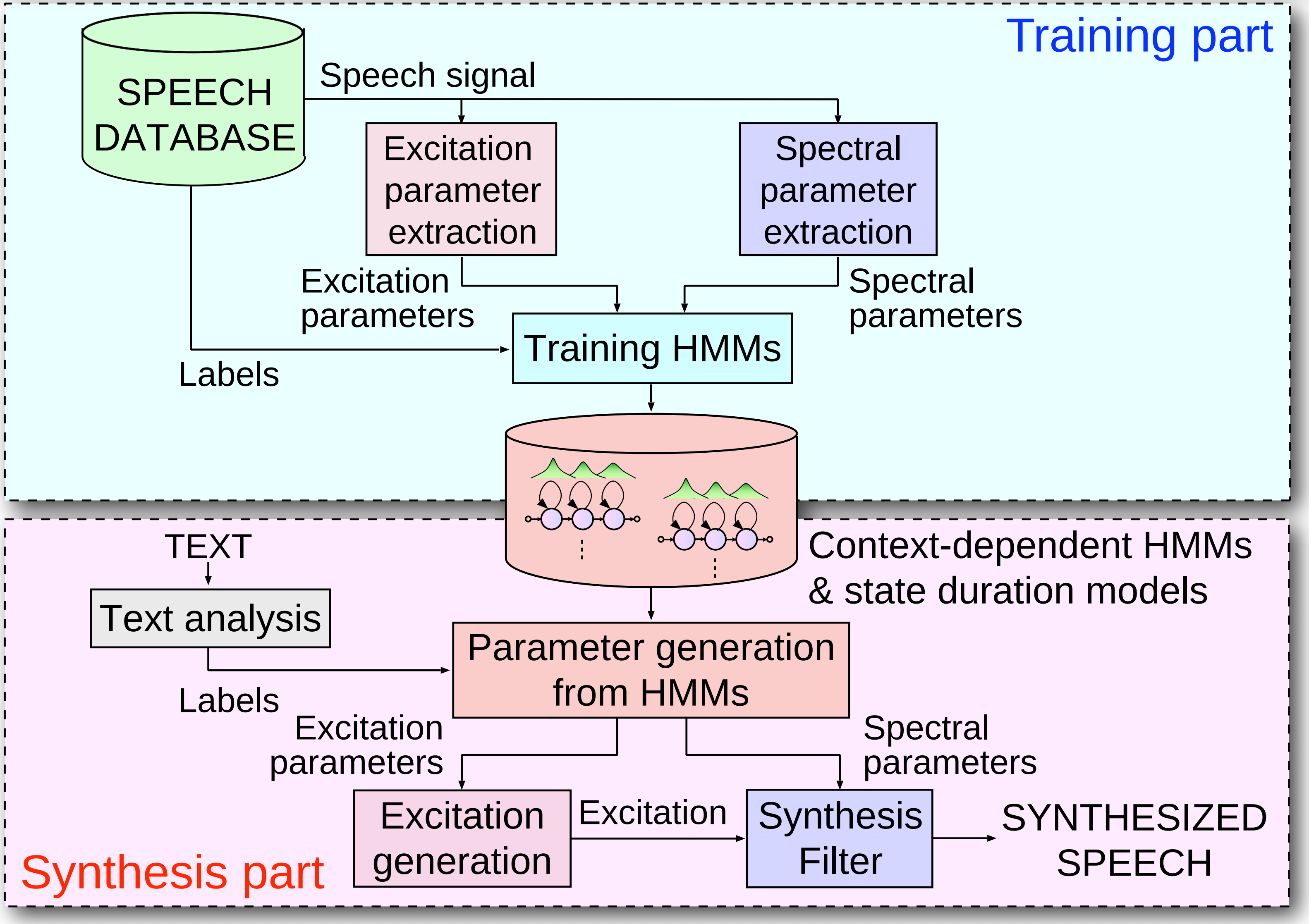
Table 1: Context labels adopted in Japanese HTS



Previous phoneme identity
Current phoneme identity
Next phoneme identity
Position of the current mora in the current accent phrase
Difference between accent type and position of the current mora
POS of the previous word
Inflected form of the previous word
Conjugation type of the previous word
POS of the current word
Inflected form of the current word
Conjugation type of the current word
POS of the next word
Inflected form of the next word
Conjugation type of the next word
Number of morae of the previous accent phrase
Accent type of the previous accent phrase
Connection intensity between the previous accent phrase and the current accent phrase
Pause existence between the previous accent phrase and the current accent phrase
Number of morae in the current accent phrase
Accent type in the current accent phrase
Connection intensity between the previous accent phrase and the next accent phrase
Position of the current accent phrase in the current breath group
Interrogative sentence or not
Number of morae of the next accent phrase
Accent type of the next accent phrase
Connection intensity between the next accent phrase and the current accent phrase
Pause existence between the next accent phrase and the current accent phrase
Number of morae of the previous breath group
Number of morae of the current breath group
Position of the current breath group in the sentence
Number of morae of the next breath group
Number of morae of the sentence

A diagram of HMM-based synthesis

HMM-based speech synthesis system (HTS)

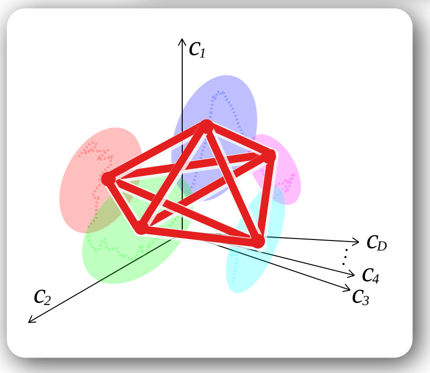


Title of each lecture

- Theme-1
  - ~~Multimedia information and humans~~
  - ~~Multimedia information and interaction between humans and machines~~
  - ~~Multimedia information used in expressive and emotional processing~~
  - ~~A wonder of sensation - synesthesia -~~

- Theme-2
  - Speech communication technology - articulatory & acoustic phonetics -
  - Speech communication technology - speech analysis -
  - Speech communication technology - speech recognition -
  - Speech communication technology - speech synthesis -

- Theme-3
  - A new framework for “human-like” speech machines #1
  - A new framework for “human-like” speech machines #2
  - A new framework for “human-like” speech machines #3
  - A new framework for “human-like” speech machines #4



Assignment

- Assignment
  - Read a research paper which is related to the second four lectures of this class, summarize it, and give your own comments to the paper.
    - Phonetics, speech science, and speech technology
  - All the materials used in the lectures can be available at:
    - http://www.gavo.t.u-tokyo.ac.jp/~mine/japanese/media2015/class.html
- Length
  - A few pages of A4 size.
- Submission
  - Your report should be sent to mine@gavo.t.u-tokyo.ac.jp in the form of PDF.
  - The file name should be “[student id]_[your name].pdf”
  - The paper that you read should be attached.
- Deadline
  - Dec. 15th